Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent

Year: 2023 | Citations: 410 | Authors: Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin

---

Abstract

Large Language Models (LLMs) have demonstrated remarkable zero-shot generalization across various language-related tasks, including search engines. However, existing work utilizes the generative ability of LLMs for Information Retrieval (IR) rather than direct passage ranking. The discrepancy between the pre-training objectives of LLMs and the ranking objective poses another challenge. In this paper, we first investigate generative LLMs such as ChatGPT and GPT-4 for relevance ranking in IR. Surprisingly, our experiments reveal that properly instructed LLMs can deliver competitive, even superior results to state-of-the-art supervised methods on popular IR benchmarks. Furthermore, to address concerns about data contamination of LLMs, we collect a new test set called NovelEval, based on the latest knowledge and aiming to verify the model's ability to rank unknown knowledge. Finally, to improve efficiency in real-world applications, we delve into the potential for distilling the ranking capabilities of ChatGPT into small specialized models using a permutation distillation scheme. Our evaluation results turn out that a distilled 440M model outperforms a 3B supervised model on the BEIR benchmark. The code to reproduce our results is available at www.github.com/sunnweiwei/RankGPT.